Statistical Methods for Data Science

Mini Project #4

Yagna Srinivasa Harsha Annadata

Yxa210024

Problem 1:

The scatter plot in Figure 1 illustrates a weak positive correlation between GPA and ACT. The sample correlation coefficient between GPA and ACT is calculated to be 0.2694818, further confirming the weak positive relationship. The bootstrap results, presented in Table 1, reveal that the estimated correlation coefficient has a bias of 0.004057848 and a standard error of 0.1049293. The confidence interval suggests that plausible values for ρ (the population correlation coefficient) fall between 0.0701 and 0.4791

Figure 1:

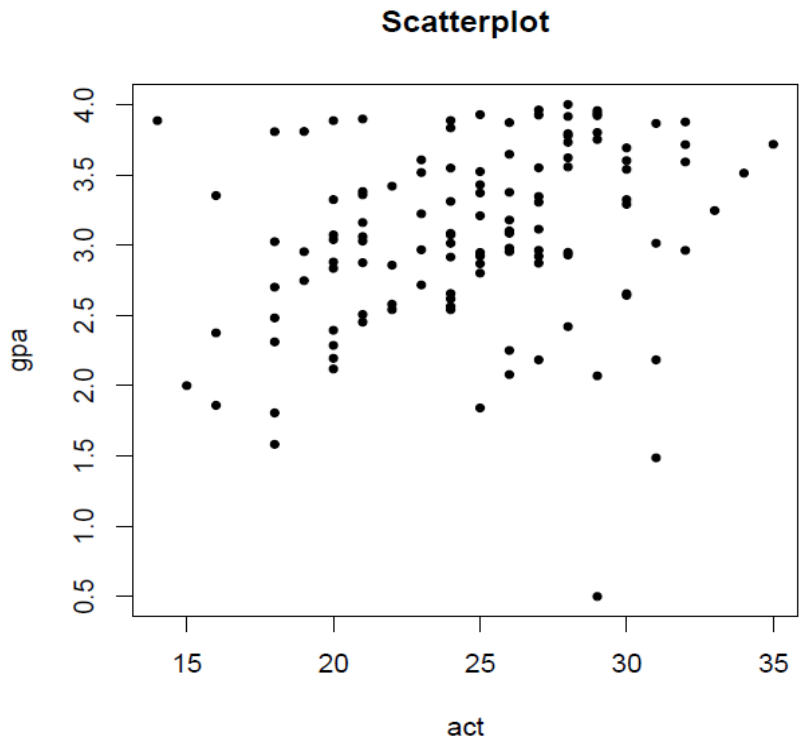

Table 1:

	ρ
Estimate	0.2694818
Bias	0.004057848
Standard Error	0.1049293
Confidence Interval min	0.0701
Confidence Interval max	0.4791

Rcode:

```
> #Load the necessary libraries
> library(boot)
Attaching package: 'boot'
The following object is masked _by_ '.GlobalEnv':
motor
> #Load the data from CSV file
> gpa <- read.csv("C://Users//yxa210024//Desktop//Masters//spring2023//Stats
for DS//mini_project
4//gpa.csv")
> #Attach the data for easy access to variables
> attach(gpa)
The following object is masked _by_ .GlobalEnv:
gpa
> #Generate scatter plots
> plot(gpa ~ act, data = gpa, pch = 20, main = "Scatterplot")
> #Calculate the Pearson correlation coefficient between gpa and act
> Cor_relation <- cor(gpa, act)
> #Define a function for bootstrap resampling
> Correlation_Resampling_function <- function(corr, indices){
+ cor(corr[indices, 1], corr[indices, 2])
+ }
> #Define a function to get the bootstrapped samples
> Corr_boots_samp <- function(corr, i = c(1:n))
+ {
+ c <- corr[i,]
+ return(c)
+ }
> #Set the number of bootstrap replicates
> Boot_replicates <- 10000
> set.seed(1000)
> #Perform bootstrap resampling for correlation between gpa and act
> Cor_boot_resample <- boot(gpa, Correlation_Resampling_function, R =
Boot_replicates)
> #Display the results of bootstrap resampling
> Cor_boot_resample
ORDINARY NONPARAMETRIC BOOTSTRAP
Call:
boot(data = gpa, statistic = Correlation_Resampling_function,
R = Boot_replicates)
Bootstrap Statistics :
original bias std. error
t1* 0.2694818 0.004057848 0.1049293
> #Calculate the 95% percentile confidence interval for correlation between
gpa and act
> boot.ci(Cor_boot_resample, conf = 0.95, type = "perc")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 10000 bootstrap replicates
CALL :
boot.ci(boot.out = Cor_boot_resample, conf = 0.95, type = "perc")
Intervals :
Level Percentile
95% ( 0.0701, 0.4791 )
Calculations and Intervals on Original Scale
```

Problem 2:

(a)

Figure 2 displays side-by-side boxplots illustrating the distribution of voltage based on location. Table 2 presents summary statistics, which reveal notable differences between the two distributions. Specifically, based on measures of location such as mean, Q1, and Q3, it appears that devices set up at remote locations tend to have higher voltage compared to those at local locations. Additionally, the distribution of voltage for local devices appears to have higher variability, as indicated by the interquartile range (IQR). Furthermore, the distribution of voltage for remote devices appears to be left-skewed, while that of local devices appears to be symmetric. Notably, there are some unusually high and low voltage readings for remote devices.

(b)

The normal quantile-quantile (Q-Q) plots for these data, presented in Figure 3, suggest that the normality assumption may hold for local devices but not for remote devices. However, despite this deviation from normality, we proceed with a t-test as the differences between the two distributions are stark and any reasonable statistical procedure would yield similar conclusions. Alternatively, we could also consider using bootstrap or nonparametric methods in this situation. We do not make any assumptions about the equality of variances, and therefore, we use the Satterthwaite approximation to calculate the confidence interval (CI). The 95% CI for the difference in means between remote and local devices is [0.1172284, 0.6454382], indicating that the mean voltage for remote devices exceeds that of local devices by an amount between 0.1172284 and 0.6454382. Thus, we can conclude that there is a statistically significant difference in the mean voltage between remote and local devices. This conclusion is supported by the results of a two-sample t-test.

(c)

As evident from Table 2 and Figure 2, the voltage distribution for remote locations is noticeably shifted to the right compared to the distribution for local locations. This is evident from the larger sample quartiles for remote locations compared to local locations, suggesting that the distribution for remote locations may have a higher mean. The results from part (b) confirm this finding.

Figure 2:

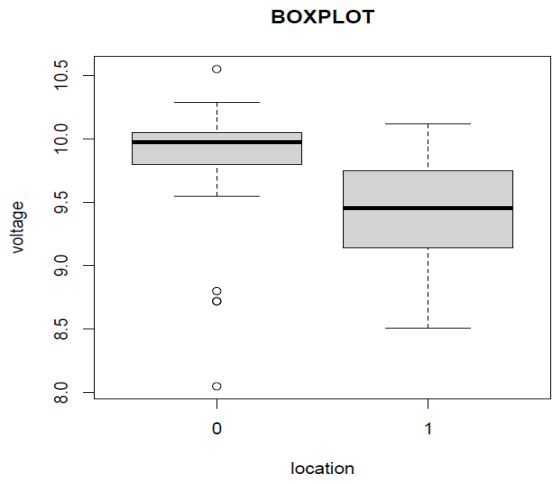

Figure 3:

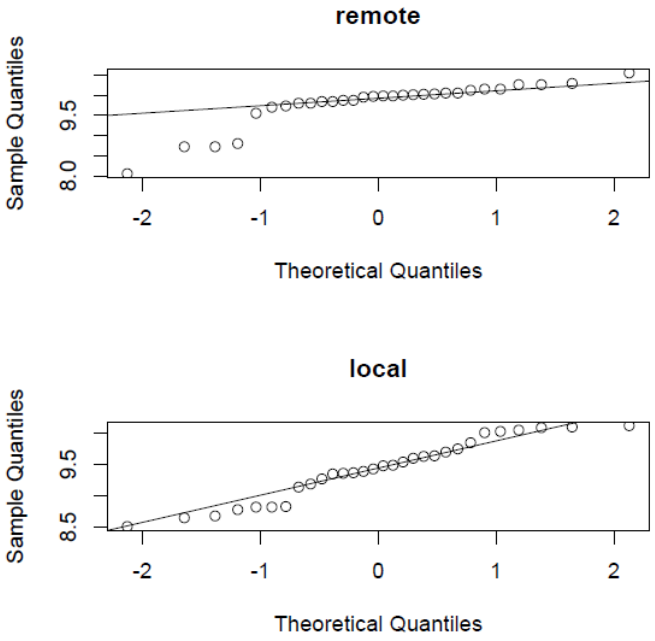

Table 2:

	Remote	Local
Min	8.0500000	8.5100000
Q1	9.8000000	9.1525000
Median	9.9750000	9.4550000
Mean	9.8036667	9.4223333
Q3	10.0500000	9.7375000
IQR	0.2500000	0.5850000
Max	10.5500000	10.1200000
SD	0.5409155	0.4788757

Rcode:

```
> #Load the data from CSV file
> volt <- read.csv("C://Users//yxa210024//Desktop//Masters//spring2023//Stats
for DS//mini_projec
t4//voltage.csv")
>
> #Attach the data for easy access to variables
> attach(volt)
>
> #Generate boxplots
> boxplot(voltage ~ location,main="BOXPLOT")
>
> #Define a custom function to calculate summary statistics
> new.summary <- function(x){
+ result <- summary(x)
```

```
+ result_summary<- c(result[-6], IQR = IQR(x), result[6], SD = sd(x))
+ return(result_summary)
+ }
>
>
> #Calculate summary statistics by location using the custom function
> by(voltage, location, new.summary)
location: 0
Min. 1st Qu. Median Mean 3rd Qu. IQR Max.
8.0500000 9.8000000 9.9750000 9.8036667 10.0500000 0.2500000 10.5500000
SD
0.5409155
------------------------------------------------------------
location: 1
Min. 1st Qu. Median Mean 3rd Qu. IQR Max.
8.5100000 9.1525000 9.4550000 9.4223333 9.7375000 0.5850000 10.1200000
SD
0.4788757
>
> #Subset the data for remote and local locations
> remote <- volt[which(location == 0), "voltage"]
> local <- volt[which(location == 1), "voltage"]
> #Generate normal QQ-plots for remote and local locations
> par(mfrow=c(2, 1))
> qqnorm(remote, main = "remote")
> qqline(remote)
> qqnorm(local, main = "local")
> qqline(local)
>
R Console Page 2
> #Calculate confidence interval using t-test
> T_test <- t.test(remote, local)
> Confidence_Interval <- T_test$conf.int
> Confidence_Interval
[1] 0.1172284 0.6454382
attr(,"conf.level")
[1] 0.95
```

Problem 3:

First, we conduct an explanatory analysis on the data, presenting the summary statistics of theoretical and experimental vapor pressure, as well as their differences, in Table 3. Additionally, Figure 4 and Figure 5 displays boxplots of the data. The quartile estimates (Q1, median, and Q3) for theoretical pressure closely resemble those of experimental pressure, indicating similar distributions. This is further supported by the summary statistics of differences.

Subsequently, we perform a paired t-test on the compound dibenzothiophene, as the data consists of paired observations (experimental and calculated) at a given temperature. The normal Q-Q plot of the difference data, shown in Figure 6, suggests that the assumption of normality is reasonable. The null hypothesis (H0) posits that the true mean difference between experimental and calculated values is zero ($\mu d = 0$), while the alternative hypothesis (H1) posits that the true mean difference is not equal to zero ($\mu d \neq 0$). The 95% confidence interval for the paired t-test is [-0.0068, 0.0083], which includes zero. Consequently, we accept the null hypothesis, indicating that the theoretical model for vapor pressure is a reliable representation of reality. This finding aligns with the results of the explanatory data analysis. Overall, our findings support the validity of the theoretical model for vapor pressure. This result is also consistent with the outcomes of the explanatory data analysis.

Figure 4:

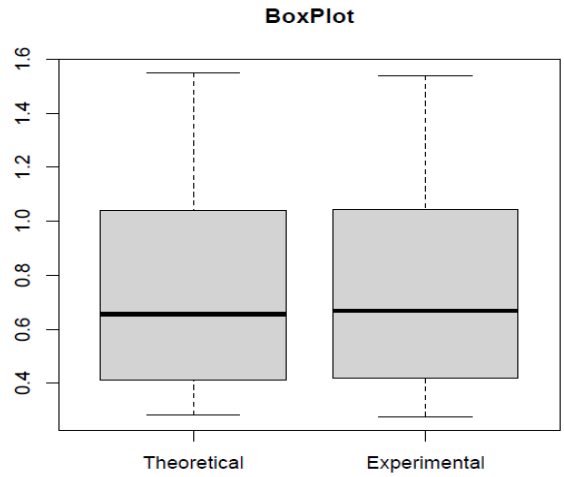

Figure 5:

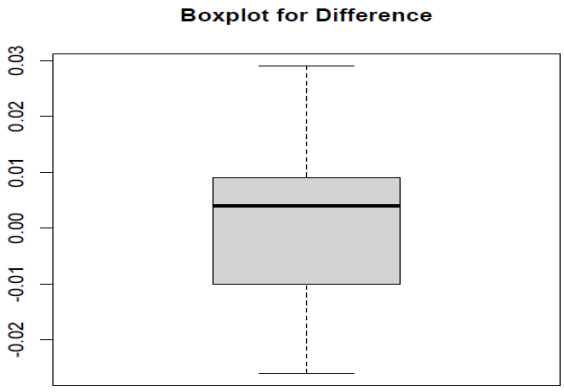

Figure 6:

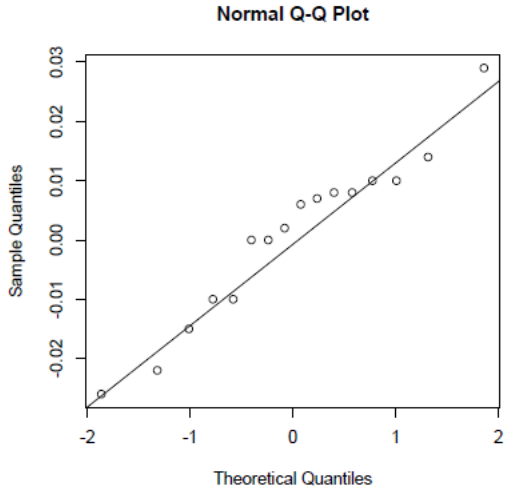

Table 3:

	Theoretical	Experimental	Difference
Min	0.2820	0.2760	-0.0260000
Q1	0.4175	0.4305	-0.0100000
Median	0.6555	0.6675	0.0040000
Mean	0.7606	0.7599	0.0006875
Q3	1.0250	1.0275	0.0085000
Max	1.5500	1.5400	0.0290000

Rcode:

```
> #Load the data from CSV file
> vapor <-
read.csv("C://Users//yxa210024//Desktop//Masters//spring2023//Stats for
DS//mini_proje
ct4//vapor.csv")
>
> #Attach the data for easy access to variables
> attach(vapor)
> difference <- theoretical - experimental
>
> #Create boxplots for theoretical and experimental data
> boxplot(theoretical, experimental, names = c("Theoretical",
"Experimental"),main="BoxPlot")
> #Create boxplot for difference data
> boxplot(difference, main = "Boxplot for Difference")
> #Create normal QQ plot for difference data
> qqnorm(difference)
> qqline(difference)
> #Generate summary statistics for theoretical, experimental, and difference
data
> summary(theoretical)
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.2820 0.4175 0.6555 0.7606 1.0250 1.5500
> summary(experimental)
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.2760 0.4305 0.6675 0.7599 1.0275 1.5400
> summary(difference)
Min. 1st Qu. Median Mean 3rd Qu. Max.
-0.0260000 -0.0100000 0.0040000 0.0006875 0.0085000 0.0290000
> #Perform paired t-test
> t.test(theoretical, experimental, paired = TRUE)
Paired t-test
data: theoretical and experimental
t = 0.19344, df = 15, p-value = 0.8492
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
-0.006887694 0.008262694
sample estimates:
mean difference
0.0006875
> #Extract confidence interval from t-test result
> conf_interval <- t.test(theoretical, experimental, paired = TRUE)$conf.int
> conf_interval
[1] -0.006887694 0.008262694
attr(,"conf.level")
[1] 0.95
```